Dark-Web Forum Analysis Using Bidirectional Encoder Representations from Transformers

Anu Varghese*
School of Computing, Wichita State University Wichita, Kansas, USA

There is a significant rise in cyber fraud, illegal activities such as stealing credit card information, and social security numbers, sale of drugs and weapons, hiring hitmen and the list goes on. The dark web has been a suitable platform for all the listed activities because of its anonymous nature. According to dark web statistics of Tor users in 2022, the US users were 34.81% of the Dark Web daily user count which translates to 831,911 users. Russia ranked second with 11.46%, followed by Germany with 7.16%. Researchers have been investigating the ways to study how these dark web marketplaces operate. This project sheds light on a marketplace called 'wall street' by analyzing the real conversations on the various subforums in it and the actors involved.

I. INTRODUCTION AND MOTIVATION

The dark web is the hidden layer of deep web consisting of hidden websites which can only be accessed with certain software, such as Tor. Since the dark web consists of IP addresses that are routable but not in use, it is difficult to track the data stream which makes it convenient for criminal activities such as illegal goods and services like selling drugs, weapons, explosives, stolen data, collecting ransom money, hiring hitmen, using botnets to launch DDOS attack on legitimate websites. People engaged in sex trafficking and child pornography leverage the anonymity of dark web. The unknown nature of dark web can be explored to get insights on how these forums and people involved in it operate.

The advent of cryptocurrencies like Bitcoin, Dash, Monero and Litecoin have made the purchase of illegal goods simple as these enable the users to trade without leaving a trace of money trail. There are many challenges to access and monitor dark web forums for many reasons like their short lifespan and changing domain addresses every now and then. For these reasons, security professionals need to always be on the lookout for newly emerging dark web markets.

The project aims at analyzing the popularity of the subforums in wall street marketplace, by considering the reviews and posts from people who sell, buy, and are involved in some way and by visualizing user engagement on each sub forums which involves implementing sentiment analysis on the reviews and posts and rating them on the scale of 0 to 4(in increasing order of user engagement and satisfaction)

This study is to create a fine-tuned model for classifying sentiments of users and to identify what category of illegal activity is the most. The data used is part of a Project initiative by the University of Arizona (NSF $\#ACI-1443019$), Drexel University, University of Virginia, University of Texas at Dallas, and University of Utah.

By examining the outcomes of this project, we can get an overall idea of how these marketplaces operate and an overview of the trending discussions on their subforums.

This report will include the literature review, design of the project, analysis, results, and future directions for this project.

II. LITERATURE REVIEW

This project's literature review includes research efforts made within the areas of classifying the dark web network traffic, identifying products and their descriptions regarding the Dark-net marketplaces and dark web activities in general, and the identified research gaps and questions. Below are the works done by various groups of researchers which contributed to the research on darknet marketplaces.

A group of researchers from Austrian Institute of Technology worked on a project where machine learning techniques were used to classify product descriptions from darknet marketplaces [1]. They worked on advertisement data (DNM archives – Grams) including offers on products and classified the items by using pre-trained word embeddings with NLP and various machine learning algorithms to classify the products. A micro-f1 score of 96% was achieved with Linear support vector machine model with TensorFlow universal sentence encoder for feature extraction.

DarkDetect: the project focused on the detection of darknet traffic by monitoring traffic from unused internet address space. The IP addresses could be spoofed or totally from unused address blocks [2]. Fine-tuned ML algorithms like Decision Tree (DT), Gradient Boosting (GB), Random Forest Regressor (RFR), and Extreme Gradient Boosting (XGB) were used along with modified Convolution-Long Short-Term Memory (CNN-LSTM) and Convolution-Gradient Recurrent Unit (CNN-GRU) to identify the network traffic accurately and results were compared.

A study on the recent shutdown drug market called 'Hydra' (April 2022) was done by a group of researchers by analyzing the drug distribution listings, and market concentration. They also looked at the total operation

* axvarghese1@shockers.wichita.edu

from listing, escrow, the delivery of drugs, and user feedback. On examining the user feedback, the ratings were high[3].

An approach to identify cyber threats in the darknet marketplace using data science was proposed in [4] where a semi-supervised labeling technique to label based on lexical and structural characteristics of darknet markets using transductive learning where they evaluated large collection of data and applied Long short-term memory for improving accuracy.

The previous works focuses on classifying and analyzing network traffic and trends in marketplaces and focused on the reviews and descriptions of the marketplace products like drugs and overview of the activities on various DNMs. This project focused more on analyzing the posts from users and form an insight based on the same by using state-of-the-art BERT model which works best for sentiment analysis.

A. Research Gaps and Questions

Gaps were identified following this literature review within the scope of Dark-net analyses. Analysis of products and description would only shed light on the activities involved in darknet markets. There have not been many studies focusing on the users involved or about the sellers. Hence, the user feedback plays important role. Based on these gaps, the following questions are formed.

- What are the contents posted by people on darknet marketplaces?
- What is the user engagement in the darknet marketplace subforum and how popular are they among the users?

III. PROJECT DESIGN

The design for this project followed five major phases: data collection, data cleaning, fine-tuning BERT, class prediction and data visualization and is given in Fig. (1).

A. Datasets

The dataset consists of postings from the “Wallstreet” darknet marketplace forum containing 45,372 rows and 14 columns including fields like post ID, thread ID, thread title, URL, subforum, author name, author’s membership, author’s join date, author’s reputation, post date, number of likes, and post content. The data was collected in the 2018-19 period. The dataset is from AZSecure-data.org which is a data science testbed for security researchers and is a project undertaken by the University of Arizona, Drexel University, the University of Virginia, the University of Texas at Dallas, and the University of Utah.

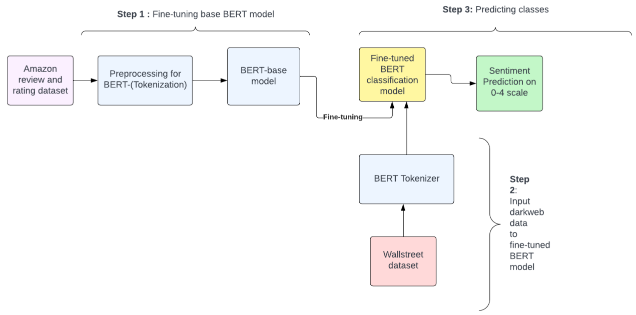

Step 1 : Fine-tuning base BERT model
Step 3: Predicting classes
Amazon review and rating dataset
Preprocessing for BERT-(Tokenization)
BERT-base model
Fine-tuning
Fine-tuned BERT classification model
Sentiment Prediction on 0-4 scale
BERT Tokenizer
Wallstreet dataset
Step 2: Input darkweb data to fine-tuned BERT model

FIG. 1. The figure shows the schematic design architecture of Project. Amazon reviews and ratings dataset is used to fine-tune BERT model which then predicts sentiment classification for wall street data.

The training dataset is from Kaggle which consists of 4515 product reviews and their ratings. This dataset was used to fine-tune BERT for classifying labels for ‘wall street’ dataset which represent a sentiment polarity from 0 to 4.

B. Data Processing and Exploratory Data Analysis

The data was cleaned by removing special characters, punctuation marks, and html tags and was given as input to the BERT tokenizer which split the data into tokens with input_ids and attention masks. The wall street dataset contains various fields which do not contribute to the problem at hand. Hence, only the user comments and corresponding subforums were taken into consideration. As a step of exploratory data analysis, we can derive an insight from the plot for the popular forums based on user engagement in the form of posts and reviews on various subforums and is given in Fig. (2).

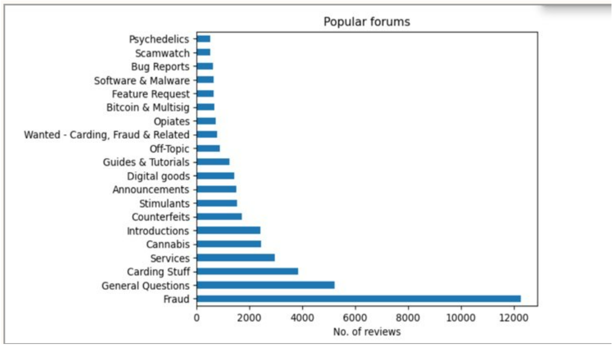


FIG. 2. Popular subforums based on number of reviews.

C. Model Design : Fine-tuning BERT

Bidirectional Encoder Representations from Transformers (BERT) is a transformer-based machine learning technique for natural language processing (NLP) developed by Google[5]. It can be fine-tuned to carry out different tasks pertaining to NLP. In this project, the BERT model was trained with Amazon reviews and rating datasets such that it can predict a user review or post based on the sentiment scale of 0 to 4, where 0 is the bad rating and 4 is the highest rating.

The model involves neural network layers which include the input_ids layer, attention mask layer, and the TFBertModel layer with global maxpooling layer which with activation function used is '*relu*' and the last dense layer is the output layer for predicting 5 labels and is given in detail in Fig. (3).

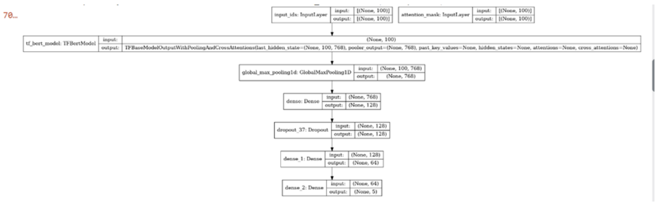
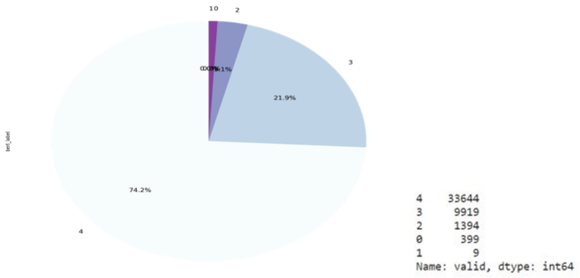

70...
input_ids: InputLayer | input: [(None, 100)] | output: [(None, 100)]
attention_mask: InputLayer | input: [(None, 100)] | output: [(None, 100)]
tf_bert_model: TFBertModel | input: (None, 100) | output: TFBaseModelOutputWithPoolingAndCrossAttentions(last_hidden_state=(None, 100, 768), pooler_output=(None, 768), past_key_values=None, hidden_states=None, attentions=None, cross_attentions=None)
global_max_pooling1d: GlobalMaxPooling1D | input: (None, 100, 768) | output: (None, 768)
dense: Dense | input: (None, 768) | output: (None, 128)
dropout_37: Dropout | input: (None, 128) | output: (None, 128)
dense_1: Dense | input: (None, 128) | output: (None, 64)
dense_2: Dense | input: (None, 64) | output: (None, 5)

FIG. 3. The figure depicts neural network layer which build the fine tuned BERT model. It takes reviews tokenized into input_ids and attention_masks as inputs and classifies the sentiment in the polarity of 0 to 4.

D. Results and Analysis

The model's highest recorded accuracy was recorded to be 92% with 5 epochs and batch size of 32, which depicts that the model was trained well and could predict the sentiment score of a user review with an accuracy of 92%.



```
4    33644
3     9919
2     1394
0      399
1        9
Name: valid, dtype: int64
```

FIG. 4. The figure shows the analysis of the wall street data set. The data is categorized from 0 to 4 where 0 represents the worst rating and 4 corresponds to the best case. This is done for making the analysis simpler.

As part of analyzing the results, data visualization was done and the results are shown in Fig. (5, 6). Where the data is studied such that the fine-tuned BERT model predicts the classification labels from 0 to 4 in the increasing order of satisfaction where 0 being the lowest and 4 being the highest. In Fig. (5), the percentage of the highest user satisfaction rating among subforums is depicted whereas in Fig. (6) the percentage of the lower user satisfaction rating among subforums is shown. The highest user engagement was found to be in the subforum called 'fraud' where fraudulent items are sold and similar activities are discussed.

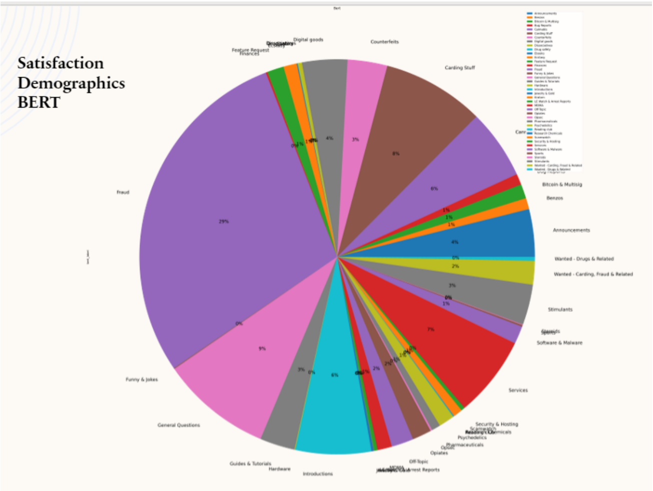


FIG. 5. The figure shows sentiment prediction by fine-tuned BERT and represents the highest user satisfaction rating in subforums.

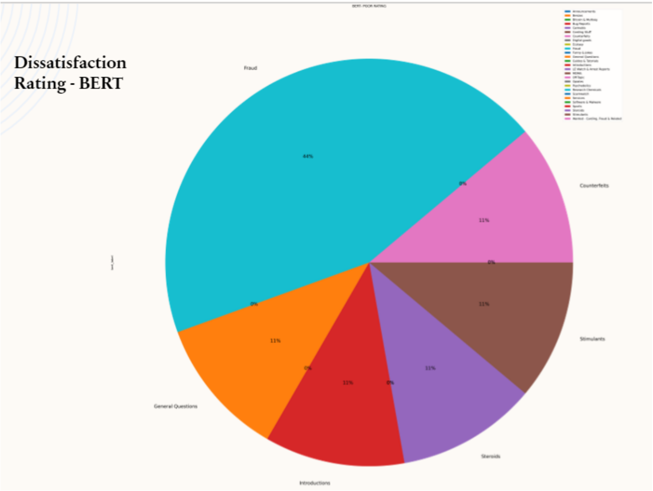


FIG. 6. The figure shows sentiment prediction by fine-tuned BERT and represents the lowest user satisfaction rating in subforums

E. Additional Work

As an extension to the proposed model, an unsupervised method to classify the wall street dataset based on sentiment analysis called the ‘Vader Algorithm’ was used. It classified the labels with a very low accuracy of 31% as compared to the fine-tuned BERT model. The main difference between BERT and Vader algorithm is that BERT considers the context of the sentence whereas Vader classifies data according to words instead of the whole sentence.

The results for the ‘Vader algorithm’ is given in Fig. (7 and 8). Similar to the previous analysis, Fig. (7)m shows the percentage of the highest user satisfaction rating among subforums, whereas in Fig. (8) the percentage of the lower user satisfaction rating among subforums is shown.

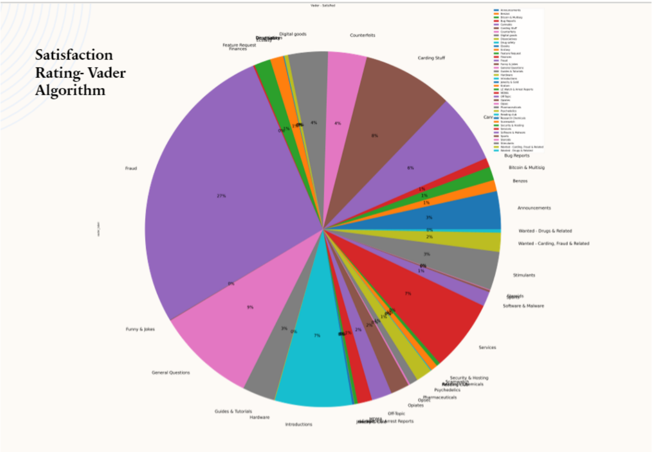


FIG. 7. This figure represents the sentiment prediction using Vader algorithm : Highest rating percentage among subforums.

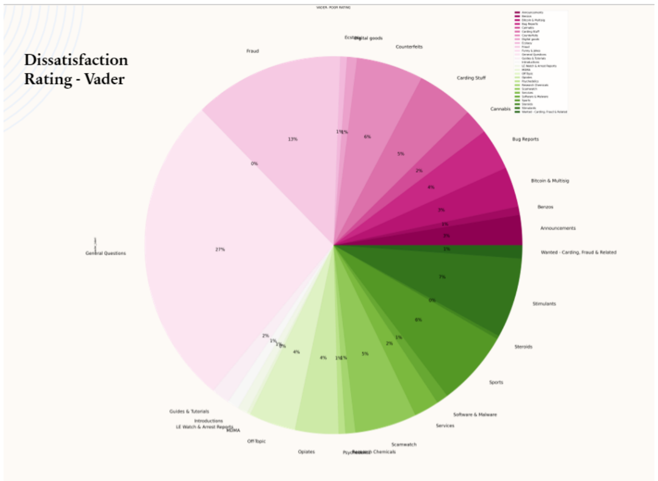


FIG. 8. This figure represents the sentiment prediction using the Vader algorithm: Lowest rating percentage among subforums.

IV. CONCLUSION AND FUTURE WORK

In this report, we studied the wall street darknet marketplace using the popular machine learning technique, natural language processing. We leveraged the state-of-the-art BERT model for sentiment classification and prediction. It yielded 92% accuracy in training and prediction. Fine-tuning the pre-trained BERT model enabled us to predict 5 classes in the increasing order of satisfaction.

The analysis helped to derive insight into user engagement and discussion on various subforums in the wall street darknet marketplace, the actors involved, the popular service/product, and the user sentiment polarity through the sentiment prediction for the corresponding subforums. This model architecture can work for a wide range of datasets to find out solutions for text classification problems based on sentiment analysis.

On analyzing the results from the Vader classification algorithm, we found out that BERT performs better as it considers the context of the sentences, and the Vader algorithm could only predict classes 31% correctly.

Future works can include training the model further to classify different specific sentiments like happiness, sadness, anger, etc instead of sentiment polarity. The dataset can be analyzed to derive more insights into user activities. Different machine learning and deep learning techniques can be used to classify and predict the sentiment for similar kinds of data.

ACKNOWLEDGMENTS

The author would like to thank the project advisor Dr. Sergio Salinas Monroy for useful discussions about the dark web, and the darknet marketplaces as well as other useful insights into this topic.

[1] Clemens Heistracher, Franck Mignet, and Sven Schlarb, "Machine learning techniques for the classification of product descriptions from darknet marketplaces." in ICAI (2020) pp. 128–137.

[2] Muhammad Bilal Sarwar, Muhammad Kashif Hanif, Ramzan Talib, Muhammad Younas, and Muhammad Umer Sarwar, "Darkdetect: darknet traffic detection and categorization using modified convolution-long short-term memory," IEEE Access **9**, 113705–113713 (2021).

[3] Priyanka Goonetilleke, Alex Knorre, and Artem Kuriksha, "Hydra: A quantitative overview of the world's largest darknet market," Available at SSRN 4161975 (2022).

[4] Mohammadreza Ebrahimi, Jay F Nunamaker Jr, and Hsinchun Chen, "Semi-supervised cyber threat identification in dark net markets: a transductive and deep learning approach," Journal of Management Information Systems **37**, 694–722 (2020).

[5] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," arXiv preprint arXiv:1810.04805 (2018).